**Instructions:**

**There are a total of six (6) multi-part questions, with point values noted for each question. You can use R/Python or Excel unless specified otherwise.**

**Please show your calculations, or the details of your program(s) for each problem. You must supply the R/Python programs, and the programs should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to CANVAS.**

**Problem 1 – C5.0 (15 points)**

The “absenteeism\_1” CSV dataset in CANVAS, shows the level of employee absenteeism (Abs\_cat=Abs\_high, Abes,med or Abs\_low). Delete all the rows with missing values. **Use C5.0** to classify the employees. What is the accuracy of your classification? What is the precision of your classification for Abs\_cat=Abs\_High

(30% test, 70% training)

**Problem 2 – Random Forest (15 points)**

The “absenteeism\_1” CSV dataset in CANVAS, shows the level of employee absenteeism (Abs\_cat=Abs\_high, Abes,med or Abs\_low). Delete all the rows with missing values. **Use Random Forest with 500 trees** to classify the employees. What is the accuracy of your classification? What is the precision of your classification for Abs\_cat=Abs\_High

(30% test, 70% training)

**Problem 3 – CART (15 points)**

The “absenteeism\_1” CSV dataset in CANVAS, shows the level of employee absenteeism (Abs\_cat=Abs\_high, Abes,med or Abs\_low). Delete all the rows with missing values. Use CART to classify the employees. What is the accuracy of your classification? What is the precision of your classification for “Abs\_cat=Abs\_High”.

(30% test, 70% training)

**Problem # 4 – Clustering (15 points)**

The “absenteeism\_0” CSV dataset in CANVAS, shows the level of employee absenteeism (Abs\_cat=Abs\_high, Abes,med or Abs\_low). Delete all the rows with missing values. **Use Hierarchical and K-means** clustering Algorithms to cluster the recordinto three clusters (do not use abs\_cat in your clustering).

Show your clusters vs. “Abs\_cat”. Show the centroid of each K-means cluster.

**Use Excel to solve the following two problems.**

**Problem # 5– ANN (20 points)**

Using the data in the table below; construct a Neural Network with one Output Layer (z) and one Hidden Layer (two nodes A and B). Calculate the predicted outcome if the inputs to the input nodes are (Node 1=.3, Node 2=.6 Node 3= .6 and Node 4=.3)

Use the actual value of .85 and a learning factor of .1 to adjust the weight for A to z and Node 1 to A.

|  |  |  |
| --- | --- | --- |
| **From** | **To** | **Weight** |
| X | A | 0.5 |
| Node 1 | A | 0.6 |
| Node 2 | A | 0.8 |
| Node 3 | A | 0.6 |
| Node 4 | A | 0.2 |
| x | B | 0.7 |
| Node 1 | B | 0.9 |
| Node 2 | B | 0.8 |
| Node 3 | B | 0.4 |
| Node 4 | B | 0.2 |
| xx | z | 0.5 |
| A | z | 0.85 |
| B | z | 0.85 |

**Problem # 6 – CART (20 points)**

Use Excel, and the **C4.5** methodology to develop a classification model/split for the “Abs\_cat” outcome using the “absenteeism\_2” CSV training dataset (one level only):

**Dataset description:**

**Absenteeism\_0**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Transportation\_expense | Numeric | Cost of employee's travel to work |
| Distance\_from\_Residence\_to\_Work | Numeric | Distance of employee’s home from work |
| Age | Numeric | Age of the employee |
| Abs\_cat | Categorical | Absenteeism category |

**Dataset description:**

**Absenteeism\_1**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Month\_of\_absence | Categorical | The month in which Employee was absent |
| Day\_of\_the\_week | Categorical | The day of the week the employee was absent |
| Social\_drinker | Categorical | Employee is a social drinker or not |
| Social\_smoker | Categorical | Employee is a smoker or not |
| Pet | Categorical | Employee is pet owner or not |
| Trans\_expense\_cat | Categorical | Cost of employee's travel to work |
| Dist\_to\_work | Categorical | Distance of employee’s home from work |
| Age\_cat | Categorical | Age category of the employee (very young, young, Middle age, higher) |
| Abs\_cat | Categorical | Absenteeism category |

**Dataset description:**

**Absenteeism\_2**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Dist\_to\_work | Categorical | Distance of employee’s home from work |
| Age\_cat | Categorical | Age category of the employee (very young, young, Middle age, higher) |
| Abs\_cat | Categorical | Absenteeism category |